Internet Archive

Written April 2016
By Alexis Rossi, Director of Media & Access, Internet Archive

The Internet Archive's mission is "universal access to all knowledge." Our goal is to build a public library that serves a global audience. This requires collecting information, storing it safely, building an infrastructure that can serve millions of people, and addressing issues of patron privacy and information access.

The organization was founded by Brewster Kahle in 1996. It is a nonprofit and is not affiliated with any government. Internet Archive serves between 2 and 3 million people every day.

We work with many partners, including libraries, museums, archives and individuals, to build our public collections. Anyone can upload media to archive.org for free to preserve it for the future, and we encourage contributors to include Creative Commons licenses with their items. Organizational funding comes from many sources, including book digitization, web archiving, donations, grants and foundations.

Web Collection

The web is ephemeral. The average lifespan of a web page is 100 days before it changes or disappears. While the web in 1996 was not as integral to our everyday lives as it is now, Brewster Kahle envisioned a day when it might be as essential a record of our lives as the daily newspaper. We must remember our past in order to learn from it, and saving web pages seemed like a fundamental step to build a global library.

The Internet Archive's first public collection was released in 1997 in association with the Smithsonian[1] and contained a selection of web pages concerning the 1996 U.S. presidential elections. We continued to expand our efforts with a web archiving pilot study with the U.S. Library of Congress in 2000[2].

[1] http://web.archive.org/web/19970126045828/http://www.archive.org/
[2] http://web.archive.org/web/20010203130300/http://archive.org/news/index.html#2000LOC

The Wayback Machine[3], which allows you to enter a URL to view archived versions of web pages, was launched in 2001[4]. The collection contained about 10 billion web resources at that time. Prior to the launch, this archive was only available to researchers and they needed to have advanced technical skills to explore it. At the time, there were people who questioned why an archive of the Internet was necessary, but 20 years later the Wayback Machine contains the only public record of the early days of this important communication medium.

We launched the Archive-It[5] service in 2005[6] to help other institutions save web resources that they thought were important. Archive-It.org provides tools that allow non-technical people to crawl and archive web resources. Over the years, it has grown to have more than 400 partners who have created thousands of highly curated collections. We also work with several national libraries around the world to complete large-scale crawls of their national domains.

As the internet has grown in size and importance, we have increased our own collection of web resources as well. Currently Internet Archive gathers about 1 billion web captures every week. The Wayback Machine has grown to about 470 billion resources.

Anyone can enter a URL into the Wayback Machine and experience past versions of web sites. The Wayback Availability API[7] allows developers to discover archived resources, and this is currently being used to fix broken links on web sites. We are also experimenting with ways to make this massive corpus easier to explore with keyword searching.

Television Collection

The Internet Archive began archiving television programs in late 2000, starting with 20 channels from the US and several other countries. At that time, streaming video online was unusual and we were not sure how to provide access to this archived content. But television is essentially an ephemeral medium, like the web – you must preserve it when it is created, or the information may be gone forever. While some entertainment programming is replayed on a regular basis, more timely material like political debates or news programming is played once or twice and then disappears from public view.

The terrorist attacks in the United States on September 11, 2001 were a global event, watched and reported around the world on television. This prompted the

---

[3] http://archive.org/web/
[4] http://web.archive.org/web/20011130142035/http://archive.org/
[5] https://archive-it.org/
[6] http://web.archive.org/web/20051124234136/http://www.archive-it.org/
[7] https://archive.org/help/wayback_api.php

Internet Archive to make a small portion of the archived television available. One week of news programming around 9/11 was made public on October 11, 2001[8]. This collection was not a memorial; it was meant to be a tool for researchers and historians. Television news reports influence people and events. We cannot research and cite our own history if we do not have access to it.

We expanded our television archiving capabilities in 2009 and began recording 60 channels from the U.S. and other countries. In 2012 we launched the Television News Archive[9]. This service allows you to search the closed captioning of selected U.S. news programs from 2009 through yesterday; we wait 24 hours to add new broadcasts to the service. The site currently contains over a million searchable broadcasts. Users see search hits for their keywords, and can play short snippets of the video surrounding their results. These snippets can be shared and cited. To watch more than these short segments, users can borrow a DVD of the full program.

The media in the Television News Archive has proven useful to researchers. Scholars have used our on-site virtual machines to analyze captions for how often geographic locations are mentioned[10] and which presidential candidates are talked about on news programs[11]. The collection has also been used to improve an audio fingerprinting tool developed at Columbia University called audfprint[12]. This technology can been used to track how video clips proliferate through news broadcasts. For example, an analysis of a political debate reveals which "sound bites" are picked up by news broadcasts and amplified for the viewing audience[13]. Software like this helps us explore the collection of archived television in new ways.

The Political Ad Archive[14] was launched in January 2016[15]. It tracks U.S. political TV ads played in primary election states and matches them up with information about funding sources and fact checking by journalists. Metadata about the ads, including funders, television markets, and how often ads were played, can be downloaded.

---

[8] https://web.archive.org/web/20011013040507/http://tvnews3.televisionarchive.org/tvarchive/html/index.html

[9] https://archive.org/details/tv

[10] http://www.theatlantic.com/technology/archive/2013/12/a-new-map-reveals-the-geography-of-american-tv-news/282443/

[11] http://www.theatlantic.com/politics/archive/2015/08/graphic-whos-the-most-popular-candidate-mentioned-on-television/402451/

[12] http://labrosa.ee.columbia.edu/matlab/audfprint/

[13] http://television.gdeltproject.org/cgi-bin/iatv_campaign2016_rdebate1/iatv_campaign2016_rdebate1_prime

[14] http://politicaladarchive.org/

[15] http://www.pbs.org/newshour/rundown/a-new-free-tool-thats-like-x-ray-glasses-for-political-ads/

Journalists have used this resource to publish many articles about the 2016 elections[16].

Video Collection

Our first video collection was released in 2001[17]. We worked with Rick Prelinger, a film collector and historian, to preserve digital copies of about 1,000 non-theatrical films from his archive. In 2001 we were not able to stream the videos online. To experience them, users had to download them -- over very slow connections -- before they could experience them. But the films still managed to find their fans, and the collections continued to grow over the years. While most of the videos are downloadable, some contributors have chosen to only allow online streaming of their media.

Today there are more than 2 million videos on archive.org and the Prelinger collection has been joined by full length feature films, live performances, cartoons, lectures, vlogs, news reels, documentaries and other genres.

Audio Collection

The first collection of audio files on archive.org was released in 2002[18]. A group called Etree[19] brought together volunteers from around the world who taped live concerts and then shared the music with each other. The group only recorded artists who had friendly taping policies. This is a tradition that began with the Grateful Dead – fans would tape concerts and then trade audiocassettes with each other. Etree had brought this tradition online, but they were constrained by storage and bandwidth limits. They could only offer a few shows at a time to their community.

Internet Archive approached Etree and offered to store all of the taped shows and make them all available, all the time, for free. They agreed to try it, and today this one community of volunteers has contributed more than 150,000 live concerts from more than 6,000 bands[20].

Another community formed around audio books. Librivox.org is a group of volunteers who create readings of public domain books. They have recorded

---

[16] http://politicaladarchive.org/press/
[17] http://web.archive.org/web/20010331221954/http://archive.org/
[18] http://web.archive.org/web/200209/http://www.archive.org/
[19] http://etree.org/
[20] http://archive.org/details/etree

thousands of books over the years, and many of them have been downloaded from archive.org millions of times[21].

Other collections of audio items soon followed, and the site today contains more than 2 million audio items including old time radio shows, podcasts, religious sermons, old 78 records, modern radio programs, and more. While most of the items can be downloaded, some of them can only be streamed by request of the creators.

Text Collection

The Internet Archive put our first collections of ebooks online in 2002[22]. To highlight some of the possibilities of these collections, we developed a Digital Bookmobile[23] – a minivan with a satellite dish for internet and equipment that allowed you to print and bind the public domain electronic books from our collection. It is cheap enough to produce the books that you can simply print a copy of the book and give it away, rather than lending it. This bookmobile has been recreated in India and Egypt to bring books to people without access to libraries.

In 2005 the Google Books project began and Internet Archive started to develop our own digitization program. We were concerned that libraries might digitize books with Google, and then make the physical books inaccessible leaving the knowledge locked up in a commercial organization. Several partners joined with us in the Open Content Alliance to build an alternative[24].

After testing many commercially available book digitization machines, we determined that we needed to build our own non-destructive scanning system. We designed a machine called the Scribe that allows an operator to take very high quality, well-lit photographs of book pages. We also developed software to turn the resulting images into attractive ebooks, including de-skewing the lines of text and cropping the pages. We use commercial optical character recognition (OCR) software to interpret the page images into text in order to create EPUBs and to allow users to search through the books. We have recently designed a more portable version of the original book digitization machine, which we call the Table Top Scribe[25].

---

[21] https://archive.org/details/librivoxaudio

[22] http://web.archive.org/web/20021001124437/http://www.archive.org/

[23] http://web.archive.org/web/20021010095421/http://webdev.archive.org/texts/bookmobile-open_house.php

[24] http://web.archive.org/web/20051007010920/http://www.opencontentalliance.org/

[25] http://blog.archive.org/2015/10/22/special-book-collections-come-online-with-the-table-top-scribe/

Our initial scanning projects focused on digitizing older books. These tend to have low circulation in libraries, and may be vulnerable to being put in storage or discarded. These digitized older books are available for download in many formats. Books can also be viewed online using open source software we developed called the bookreader[26].

Open Library[27], launched in 2008[28], was intended to have one web page for every book ever published. We pulled together metadata records from many sources and linked to electronic versions of the books when possible. OpenLibrary.org now has more than 20 million book records that are accessed by 250,000 users per day.

Eventually we began to digitize more modern books. In 2010 we began to make these modern books available on Open Library to the print disabled[29] using encrypted DAISY files. Users must register with the U.S. National Library Service to receive a decryption key that allows them to listen to the books.

Subsequently we developed a lending program to make books available to a wider range of users[30]. When we digitize a modern book, we put the book away in our physical archive and make the digital copy available to be borrowed by one person at a time. Users can borrow up to 5 books at once for a period of 2 weeks. If the user wants to read the book online, they can borrow it through our bookreader software. If they want to download a copy, they must use Adobe Digital Editions software that keeps the files secure.

The book digitization program has grown significantly. Over the years we have worked with many libraries to make digital copies of their texts. Currently we digitize about 1,000 books per day in 30 scanning centers on 5 continents. The archive.org texts collections contain about 4 million books, plus another 4 million texts like journal articles, government documents, court cases, and census records.

Image Collection

The Internet Archive partnered with NASA in 2008 to gather the digital images from many centers and put them in a collection. Video, audio and texts were also included, and a special purpose portal was built[31]. We launched the NASA images collection with about 100,000 images[32].

---

[26] https://github.com/openlibrary/bookreader
[27] https://openlibrary.org/
[28] http://web.archive.org/web/20080505074433/http://www.openlibrary.org/?
[29] http://blog.archive.org/2010/11/26/3424/
[30] http://blog.archive.org/2011/06/25/in-library-ebook-lending-program-expands-to-1000-libraries/
[31] http://web.archive.org/web/20080725062000/http://nasaimages.org/
[32] https://archive.org/details/nasa

Over the years we have also received collections of maps and images of artwork from museums. Today the image collections contain more than a million items[33].

Software Collection

The first collection of software appeared on archive.org in 2002[34]. We added to these collections over the years, but the items were not easy to use or access. Software is written to run in a specific environment; if you cannot recreate that environment, you cannot experience the software. In other words, we could pull the 1s and 0s off a game cartridge that was designed to run on a specific game console built in the 1980s, but we could not play the game.

In 2010 the Internet Archive hired its first dedicated software archivist. This curator, Jason Scott, helped to build the collection of preserved software and related media, but he also tackled the issue of access. A group of volunteers were building a system called JSMESS[35] that allows you to emulate old operating systems in a web browser. Scott worked with these volunteers to bring this technology to the archive.org site.

In 2013[36] we released the first experimental collection of historical software that could be experienced in your browser via emulation. Over the years, the software has continued to improve and additional operating systems were supported, and today[37] millions of people on the Internet Archive are experiencing historic software that was inaccessible for decades.

Physical Archive

As physical materials are transferred to digital formats, we face a decision about what to do with the physical media. There are good reasons to preserve them. They are the original, authoritative objects; if there are questions or problems with a digital version, the physical object can be used as a reference in the future. The physical object also provides redundancy in case the digital copies are damaged.

---

[33] https://archive.org/details/image
[34] http://web.archive.org/web/20021001124437/http://www.archive.org/
[35] https://github.com/jsmess
[36] http://blog.archive.org/2013/10/25/microcomputer-software-lives-again-this-time-in-your-browser/
[37] https://archive.org/details/software

Our physical archive was established in 2011[38], allowing us to preserve millions of physical books, movies, LPs, CDs, software, films and other physical media. We are trying to preserve one copy of every piece of media we are able to acquire. We then digitize the object, use the digital copy for access and store the physical item. The physical archive is designed for long-term preservation of materials, not for daily access; daily access is done through the digital copies. Because we do not provide access to the physical collections we are able to store the physical media very densely, and therefore at a lower cost than many library storage facilities.

Digital Data Preservation

The Internet Archive currently contains 25 petabytes of unique data. We store all of the data on our own servers, and everything is replicated at least twice in different physical locations. We maintain data centers at our headquarters building in San Francisco and at our physical archive buildings in Richmond, California. In addition there are partial copies of archive data in Amsterdam and at the Library of Alexandria in Egypt.

There are many challenges to storing this amount of data. We have tens of thousands of hard drives, so there is a constant flow of drives failing that need to be replaced quickly. We do audits of the files to make sure we aren't suffering from bit rot. Electricity and bandwidth costs must be met.

But the biggest challenge is keeping media accessible to the public over long periods of time as access methods change. When new browsers, tablets or phones come on the market, file formats can go out of date quickly. For example, when the original iPhone was released, the versions of videos that we used for click and play access on the site were not compatible; suddenly a large portion of our users could not play the videos in our collections. We used the original video files to regenerate millions of access versions in a new format. Maintaining accessibility requires vigilance, strong technical capabilities, and the compute power to complete these large-scale changes. As the archive of digital media objects ages, we will need to regenerate access versions many, many times over the decades and centuries.

Privacy

As a library, we are very concerned about protecting the privacy of our readers. The archive.org site is offered in https and we do not keep user IP addresses in our web logs. We believe that reader privacy is essential to ensure freedom of thought and speech.

---

38 http://blog.archive.org/2011/06/06/why-preserve-books-the-new-physical-archive-of-the-internet-archive/

The Internet Archive actively works with organizations like the Electronic Frontier Foundation (EFF) to encourage laws that protect users, and oppose laws that threaten user privacy or our ability to maintain a public library on the Internet. Over the years, we have participated in issues regarding orphan works[39], privacy[40], and laws that would harm safe harbor status[41] for sites that accept user uploads. We also successfully fought a national security letter in 2008[42] with the help of EFF and the American Civil Liberties Union (ACLU) that would have required us to divulge information about a user.

While we avoid collecting personal information about our users, we do have anecdotal information about them because they write to us, ask us questions and interact with us on social media. University reference librarians use archive.org to provide primary source material for their students. UX designers use the Wayback Machine to examine changing interface trends on the net. We have a large community who listen to our live concert archive and debate each other on our forums about the best versions of songs from their favorite bands. Families have written to thank us for saving a deceased relative's web site.

People go to the library for all kinds of reasons - educational, personal, business, entertainment - and the examples we have of people who use archive.org are just as varied. But privacy is an essential aspect of any library, and we strive to maintain that ideal in the digital world.

Access

Society evolves because of information; everything we learn, invent or create is built upon the work of others. The internet gives us the opportunity to provide equal access to information to every single person on earth regardless of income, education level, or location. Anyone with a cell phone should be able to visit a world-class library.

In this digital age, when everything is expected to be online, we need to make sure the best resources are available. We have centuries of valuable information stored in physical libraries, archives and personal collections, and we need to make sure all of it is online and accessible. Many information professionals have spent their lives

---

[39] http://blog.archive.org/2006/11/21/orphan-works-trial-nov-13th-san-francisco/

[40] http://blog.archive.org/2014/03/11/archive-and-ala-brief-filed-in-warrantless-cell-phone-search-case/

[41] http://blog.archive.org/2016/03/22/save-our-safe-harbor-submission-to-copyright-office-on-the-dmca-safe-harbor-for-user-contributions/

[42] http://blog.archive.org/2008/05/16/fbi-gag-order-against-the-internet-archive-is-rescinded/

building stunning collections of knowledge, and now they have the opportunity to share the fruits of this labor with the entire world.

Access requires that we digitize physical media, maintain copies of already-digital materials, create or improve methods for finding information, and make media available to the public.

Every library or archive in the world will eventually need to answer questions about access. How do we continue to do our jobs in a digital world? How do we make media available to patrons in ways that respect an ecosystem that includes creators, publishers, libraries and consumers? We have discussed above some of the methods the Internet Archive uses to make media available online to our patrons; some of the media in our archive can be freely downloaded, some can only be streamed, some may be borrowed in limited quantities, some can only be consumed in short snippets, some is only available to print disabled people. We give researchers access to some media in bulk via virtual machines in our data center to allow the examination and manipulation of data.

A library fulfills its mission by collecting media and making it available to the public. A digital library must be able to do the same.